Received: 24 January 2020 | Revised: 30 April 2020 | Accepted: 4 May 2020

DOI: 10.1111/eva.12998

Evolutionary Applications Open Access WILEY

ORIGINAL ARTICLE

Gene flow as a simple cause for an excess of high-frequency-derived alleles

Nina Marchi[1,2] | Laurent Excoffier[1,2]

[1]CMPG, Institute of Ecology and Evolution, University of Berne, Berne, Switzerland

[2]Swiss Institute of Bioinformatics, Lausanne, Switzerland

Correspondence
Nina Marchi, CMPG, Institute of Ecology and Evolution, University of Berne, 3012 Berne, Switzerland.
Email: nina.marchi@iee.unibe.ch

Funding information
Swiss National Science Foundation, Grant/Award Number: 310030B_166605; University of Berne: SELF2018_04

Abstract

Most human populations exhibit an excess of high-frequency variants, leading to a U-shaped site-frequency spectrum (uSFS). This pattern has been generally interpreted as a signature of ongoing episodes of positive selection, or as evidence for a mis-assignment of ancestral/derived allelic states, but uSFS has also been observed in populations receiving gene flow from a ghost population, in structured populations, or after range expansions. In order to better explain the prevalence of high-frequency variants in humans and other populations, we describe here which patterns of gene flow and population demography can lead to uSFS by using extensive coalescent simulations. We find that uSFS can often be observed in a population if gene flow brings a few ancestral alleles from a well-differentiated population. Gene flow can either consist in single pulses of admixture or continuous immigration, but different demographic conditions are necessary to observe uSFS in these two scenarios. Indeed, an extremely low and recent gene flow is required in the case of single admixture events, while with continuous immigration, uSFS occurs only if gene flow started recently at a high rate or if it lasted for a long time at a low rate. Overall, we find that a neutral uSFS occurs under more restrictive conditions in populations having received single pulses of gene flow than in populations exposed to continuous gene flow. We also show that the uSFS observed in human populations from the 1000 Genomes Project can easily be explained by gene flow from surrounding populations without requiring past episodes of positive selection. These results imply that uSFS should be common in non-isolated populations, such as most wild or domesticated plants and animals.

KEYWORDS

computer simulation, demographic analysis, gene flow, human genetics, human genome, natural selection, neutral evolution

1 | INTRODUCTION

Allele frequency changes are driven by the combined action of different evolutionary forces such as mutation, selection, genetic drift and migration (Wright, 1931), but most of the variant frequencies are expected to be rare (Ewens, 1972), leading to a L-shaped site-frequency spectrum (SFS) (Fu, 1995). However, an unexpectedly large proportion of high-frequency-derived alleles, resulting in U-shaped SFS (uSFS), has been documented in multiple species, including wild and domesticated plants (Liu, Zhou, Morrell, Gaut, & Ge, 2017; Morton, Dar, & Wright, 2009; Price et al., 2018), animals (Cooper, Burrus, Ji, Hahn, & Montooth, 2015; de Manuel et al., 2016; Murray, Huerta-Sanchez, Casey, & Bradley, 2010) and even human populations (Henn et al., 2015; Pouyet, Aeschbacher, Thiéry, & Excoffier, 2018).

Several explanations for these uSFS have been proposed. This phenomenon has been notably interpreted as a signature of positive selection at several loci (Akashi & Schaeffer, 1997; Bustamante, Wakeley, Sawyer, & Hartl, 2001), as neutral variants hitchhiking with beneficial mutations during selective sweeps would also be observed at high frequencies (Andolfatto & Przeworski, 2001; Fay & Wu, 2000; Kim & Stephan, 2000, 2002; Lapierre, Blin, Lambert, Achaz, & Rocha, 2016; Pavlidis, Jensen, & Stephan, 2010; Stephan, 2016), thus leading to a uSFS (Hahn, 2018; Ronen, Udpa, Halperin, & Bafna, 2013). This phenomenon could even be accentuated by selection fluctuating over time (Huerta-Sanchez, Durrett, & Bustamante, 2008; Przeworski, 2002). Alternatively, low-frequency-derived alleles mistakenly annotated as ancestral would lead to the emergence of high-frequency-derived variants and also create a uSFS (Baudry & Depaulis, 2003; Hernandez, Williamson, & Bustamante, 2007). uSFS can also emerge in multiple-merger coalescent models that have been developed to account for strong selective sweeps or a very large variance in reproductive success among individuals of a population (Eldon, Birkner, Blath, & Freund, 2015; Rice, Novembre, & Desai, 2018; Sargsyan & Wakeley, 2008; Tellier & Lemaire, 2014), which is not well accounted for in the classical Kingman coalescent framework. Finally, uSFS has also been shown to arise in non-isolated populations, for example in range expanding populations (Sousa, Peischl, & Excoffier, 2014), in structured populations analysed as single populations (Cutter, 2019; Lapierre et al., 2016; Wakeley, 2000) or in structured population receiving low levels of gene flow from surrounding demes (Garrigan & Hammer, 2006; Wakeley & Aliacar, 2001).

Even though most animal and plant populations are not completely isolated and receive migrants from surrounding populations, gene flow has been rarely proposed as an explanation for uSFS, and hypotheses of selection or ancestral allele mis-assignment have been preferred (Li et al., 2012; Liu et al., 2017; Qanbari & Simianer, 2014; Sabeti, 2006). However, Pouyet et al. (2018) recently showed that the uSFS observed in human populations could not be recovered under a complex demographic scenario involving an isolated population, but could be perfectly modelled under a scenario involving gene flow from an unspecified source (i.e. a ghost population), and this, in absence of any positive selection or mis-assignment of ancestral alleles. In order to better investigate the conditions leading to uSFS in non-isolated populations, we have used simulations to explore the impact of gene flow duration, onset and intensity, as well as of population size and divergence time, on the probability of observing a uSFS. Even though more complex scenarios could certainly lead to uSFS, we have simulated here two simple demographic models of *isolation with admixture* and of *isolation with immigration*, which are often used as basic population genetic models (Geneva & Garrigan, 2010; Hahn, 2018; Patterson et al., 2012; Sousa & Hey, 2013) and represent the two ends of the gene flow spectrum. We have then compared the likelihoods of these models for ten populations from the 1,000 Genomes panel where uSFS is observed.

2 | MATERIAL AND METHODS

2.1 | Simulated scenarios

We modelled a population of n haploid individuals and effective size N that receives gene flow from an unsampled and often referred to as a "ghost" population (Beerli, 2004; Slatkin, 2005), after their divergence T_{DIV} generations ago (or expressed in $2N$ units as $\tau_{DIV} = T_{DIV}/(2N)$ the parameters and their ranges are described in Table 1). However, rather than being simply a non-sampled population, this ghost population is introduced here as a convenient way to partition sampled lineages into two structured components between which coalescent events will not immediately occur. This type of partitioning is for instance found in metapopulation models with migration, where coalescent events occur rapidly during a scattering phase and more slowly during the collecting phase (Wakeley, 1999; Wakeley & Aliacar, 2001). For sake of simplicity, we tested two models at the ends of the gene flow spectrum (Figure 1): one of *isolation with admixture* (IA) and one of *isolation with immigration* (*II*). In the *IA* model, a single admixture event (i.e. a single pulse of gene flow) occurred T_{ADM} generations ago (τ_{ADM} in $2N$ units), with an admixture rate a. In the *II* model, continuous gene flow occurring at rate m per generation started T_{GF} generations ago.

2.2 | Simulated genetic data

We used the software *fastsimcoal2* (Excoffier, Dupanloup, Huerta-Sánchez, Sousa, & Foll, 2013) to simulate the genomic diversity in 100 Mb of DNA (which roughly corresponds to the neutral portion of the human genome found in Pouyet et al. (2018)) under the scenarios defined in the previous section. The simulated 100 Mb was modelled as 10,000 blocks of 1,000 independent non-recombining regions of 100 bp. Note that the SNPs simulated in this way are essentially independent (unlinked) SNPs, and that it would have been possible to simulate partially linked SNPs but more simulations would have been necessary to get the same expected SFS (Pouyet et al., 2018). However, we have performed a limited set of simulation using partially linked SNPs, to verify that our conclusions would not change if we were explicitly simulating linkage and recombination (Supporting Information 2).

TABLE 1 Parameters description and ranges used for the simulation of different scenarios

Parameter description	*IA* model	*II* model
Effective size (haploid number, *Ns*)	{4,000; 40,000}	{4,000; 40,000}
Divergence time $\tau_{DIV} = T_{DIV}/(2N)$	{0.005; 0.05; 0.25; 0.5; 2.5}	2.5
Time of single admixture event $\tau_{ADM} = T_{ADM}/(2N)$	{0; 0.025; 0.05; 0.125; 0.25}	
Onset of gene flow (T_{GF})		T_{DIV}/ {10,000; 1,000; 100; 10; 1}
Haploid sample size (*n*)	{10; 50}	10
Admixture rate (*a*)	[0:0.5]	
Immigration rate (*Nm*)		[0.01:10]

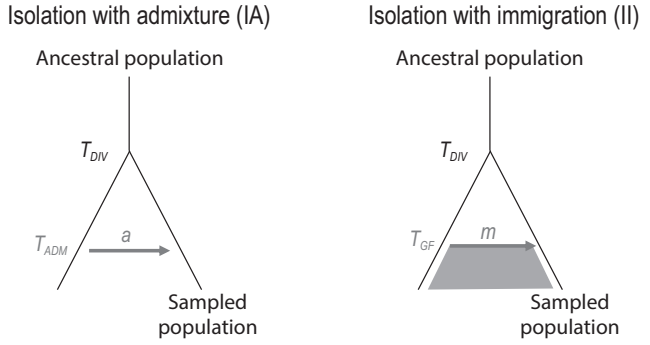

FIGURE 1 Scenarios used to elucidate conditions under which gene flow leads to a uSFS. The populations have diverged T_{DIV} generations ago from an ancestral population; their population sizes (*N*) are identical and constant over time. The results shown in the main text were obtained for a sampling size of 10 individuals and population sizes *N* = 4,000, but similar results were seen for a ten-time larger size (*N* = 40,000) after appropriate rescaling of divergence time and migration rate (Supporting Information 1)

We then computed the site-frequency spectrum (SFS) for each block independently using the *fastsimcoal2* command./fsc2 –i *File*. *par* -n 10,000 -q -c0 -d -s0 -x -l (Supporting Information 3). The mutation rate was set to 1.20×10^{-8} per bp per generation (de Manuel et al., 2016; Venn et al., 2014), and we assumed an infinite-site model. We then sampled with replacement 10,000 blocks from the original simulated set to generate a given block-bootstrap data set, and we repeated this procedure 1,000 times to generate 1,000 block-bootstrap SFS.

2.3 | Summary statistics

We computed the global unfolded SFS for each simulated and block-bootstrapped data set of 100 Mb, by summing the 10,000 (respectively, observed or randomly sampled) block-SFS. The 95% confidence intervals of the simulated SFS were computed from the 2.5% and 97.5% quantiles of the SFS entries (all SFS is shown in Supporting Information 4).

We classified simulated SFS into three categories according to their shapes: a monotonously decreasing SFS with a mode at singletons corresponding to a L-shape SFS; a U-shape SFS with a second mode at high derived allele frequencies; a W-shape SFS with a second mode at intermediate derived allele frequencies (Figure 2).

We also used a summary statistic called *D-tail* defined as.

$D-tail = \frac{SFS_{n-1} - SFS_{n-2}}{SFS_{n-2}}$, where *n* is the haploid sample size and SFS_i is the number of sites with a derived frequency *i*. *D-tail* is positive when $SFS_{n-1} > SFS_{n-2}$ (i.e. for uSFS) and negative for L-shaped and W-shaped SFS.

2.4 | Human data sets and likelihood estimations

We computed the SFS and *D-tail* statistic for ten 1,000 Genomes (1000G) populations: Yoruba in Ibadan, Nigeria (YRI); Luhya in Webuye, Kenya (LWK); Iberian Population in Spain (IBS); British in England and Scotland (GBR); Punjabi from Lahore, Pakistan (PJL); Bengali from Bangladesh (BEB); Kinh in Ho Chi Minh City, Vietnam (KHV); Japanese in Tokyo, Japan (JPT); Colombians from Medellin, Colombia (CLM); and Peruvians from Lima, Peru (PEL). We included the 10 individuals with the highest coverage per population (see Supporting table from Pouyet et al. (2018)). In this data set, we focused on the 493,369 sites formerly identified as evolving neutrally in Pouyet et al. (2018) (i.e. biallelic sites, non-CpG sites, mutations neither affected by biased gene conversion nor by background selection). The ancestral state was defined based on the chimpanzee reference genome (*panTro4*) to prevent mis-assignment of the ancestral/derived states. We used a block-bootstrap approach based on sets of 100 adjacent SNPs along the genome, to generate 1,000 block-bootstrap SFS and *D-tail* statistics.

We estimated with *fastsimcoal2* the likelihood of four demographic scenarios (Supporting Information 5) inspired from Pouyet

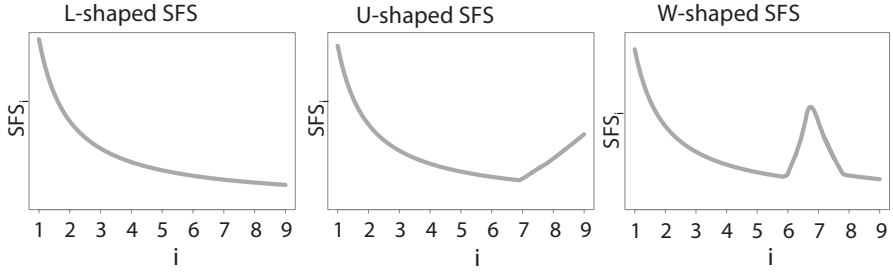

FIGURE 2 Schematic SFS shapes, for a sample of haploid size 10 where SFS_i is the number of sites with a derived frequency *i*

et al. (2018): (a) a first simple scenario, where a fully isolated population can go through four different epochs with four different sizes separated by three bottlenecks of arbitrary size and times, (b) same as the first scenario but allowing for potential ancestral-state mis-assignment (option -ASM in *fastsimcoal*), (c) same as the first scenario but allowing for continuous gene flow from a ghost population and (d) same as the first scenario but allowing for a single pulse of admixture from a ghost population. Parameters were estimated for each model with the *fastsimcoal2* command line options: -t *POP*.tpl -e *POP*.est -n200000 -d -M -L40 -q -0 -C1 -c1 -B1, where *POP* is the acronym of each of the ten 1000G populations (generic input files made available in Supporting Information 6). In order to scale parameters, we assumed an ancestral human population size of 20,000 diploids, and a constant and uniform mutation rate of 1.25×10^{-8} per bp per generation (Scally & Durbin, 2012), which is widely used in demographic inference in humans (Malaspinas et al., 2016; Pagani et al., 2015; Raghavan et al., 2015; Schiffels & Durbin, 2014; Sikora et al., 2017, 2019; Spence & Song, 2019; Steinrücken, Kamm, Spence, & Song, 2019).

3 | RESULTS

3.1 | Isolation with admixture

To evaluate the impact of the divergence time, we have first simulated an *isolation with admixture* (IA) model where the admixture event occurred at sampling time (0 generations ago) for varying divergence times (T_{DIV}) and admixture rates (a).

As expected, without admixture ($a = 0$), the SFS is L-shaped (Figure 3a and Supporting Information 7) and the D-tail statistics are negative (Figure 3b). This is also the case when $a > 0$ for recent divergence times ($\tau_{DIV} = T_{DIV}/(2N) = 0.005$ or $\tau_{DIV} = 0.05$). However, for older divergence times, when $\tau_{DIV} > 0.05$, the pattern is more complex: positive *D-tail* statistics and consistent uSFS are only observed for relatively low admixture rates (between 5% and 20%). Importantly, the admixture rates leading to uSFS depend on the sample size n: for a larger sample size ($n = 50$), we observe uSFS for reduced admixture rates ($0 < a \leq 0.03$), while larger admixture rates lead to W-shaped SFS with not only one but two internal maxima (Supporting Information 8). In any case, independently of sample sizes, *D-tail* values increase for older divergence times, indicating that SFS is more strongly U-shaped with larger divergence times.

These results are best explained by the immigration of a few ancestral alleles into the sampled population at sites where the derived allele is fixed in the sample before admixture, thus causing a decrease in the frequency of derived alleles from n to n-1. For the same amount of admixture, this phenomenon is more likely if two populations have fixed different alleles, the probability of which increases with divergence times, and becomes substantial when $\tau_{DIV} \geq 0.5$ (Hudson & Coyne, 2002). To substantiate this explanation, we have performed simulations for $\tau_{DIV} = 2.5$, where we computed derived allele frequencies after the admixture event at sites that were fixed-derived before admixture (Figure 3c). For relatively low admixture rates ($a = 0.05$), almost 80% of the previously fixed derived sites are transformed into nearly fixed sites and SFS becomes U-shaped. This proportion drops to 60% when $a = 0.1$. For larger admixture rates ($a \geq 0.2$), SFS becomes W-shaped (Supporting Information 7), as admixture events will often introduce more than one ancestral allele at previously fixed sites.

Note that under the *IA* model, large admixture rates corresponding to partial genetic replacement ($0.5 < a < 1$) can also lead to uSFS. Indeed, uSFS is also obtained for admixture rates between 0.8 and 0.95, in a way symmetrical to low a values ($0.05 < a < 0.2$, Supporting Information 9A). In this case, the excess of high derived frequencies is caused by the immigration of a large number of derived alleles at sites where the ancestral allele was fixed in the sampled population

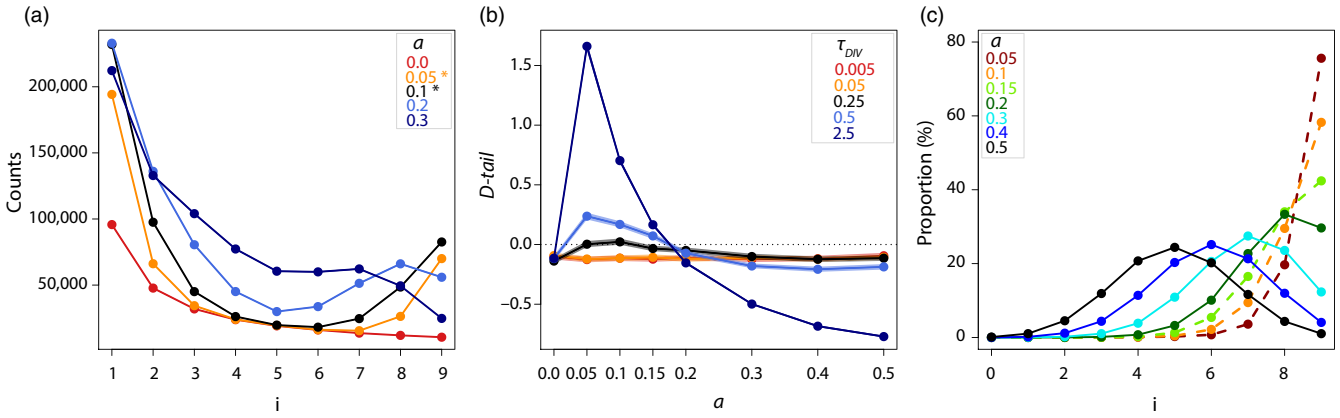

FIGURE 3 Effect of the admixture rate and time of divergence on SFS properties, under an *IA* scenario for $\tau_{ADM} = 0$. (a) SFS, i.e. the number of sites with a derived frequency i, from $n = 10$ haploid individuals for $\tau_{DIV} = 2.5$ and various admixture rates a; (b) *D-tail* statistic for various divergence times τ_{DIV}; (c) proportion of loci in the sampled population that were fixed for the derived allele before the admixture event and which show i derived alleles afterwards, when $\tau_{DIV} = 2.5$. In panes a and b, dots and solid lines were obtained from simulated data sets, and semi-transparent colours define 95% block-bootstrap confidence intervals. Note that these confidence intervals are so small that they are barely visible on these figures. In pane c, dashed lines stand for uSFS and solid lines stand for W-shaped SFS

(Supporting Information 9B), mimicking the action of positive selection (Hahn, 2018).

3.2 | Effects of the onset time of instantaneous and continuous gene flow

When gene flow occurred more than one generation ago, its onset time, intensity and duration might also have a drastic effect on the allele frequency distribution, and thus on the shape of the SFS. To investigate the effect of past gene flow, we have run simulations under both an *isolation with admixture* (IA) and an isolation with immigration (*II*) models, where the populations have diverged for $\tau_{DIV} = 2.5$ (i.e. 20,000 generations for effective populations size N = 4,000), when $\tau_{ADM} > 0$ and when $T_{GF} > 0$ for the *IA* and *II* models, respectively.

Under the *IA* model, uSFS and positive *D-tail* values are observed for admixture $\tau_{ADM} < 0.25$. SFS becomes less U-shaped, and *D-tail* values are smaller for older admixture times (Figure 4a). However, even though *D-tail* statistics are lower when admixture is old (i.e. $\tau_{ADM} = 0.125$), uSFS is observed for larger admixture rates than when it is very recent (i.e. when $\tau_{ADM} = 0$). As expected, the SFS can become multimodal for recent divergence times and large admixture rates ($\tau_{ADM} \leq 0.025$ and $a \geq 0.2$), and the internal mode moves towards more central values for larger admixture rates (Figure 4c).

Under the *II* model, we observe uSFS and positive *D-tail* statistics for a large range of onset times for gene flow (from $T_{GF} = T_{DIV}$ to $T_{GF} = T_{DIV}/10{,}000$; Figure 4b and d), but the amount of gene flow required to produce uSFS is inversely correlated with the age of the onset of gene flow; that *is*, small immigration rates ($Nm < 0.5$) are necessary when gene flow is ancient ($T_{GF} = T_{DIV}$) and large immigration rates ($Nm > 1$) are necessary when gene flow is very recent ($T_{GF} = T_{DIV}/10{,}000$) (Figure 4d). We found an exception for $T_{GF} = T_{DIV}/10$, where both low and high rates lead to uSFS, likely due to the introduction of both ancestral and derived alleles in the population, depending on which allele was fixed ancestral or fixed derived in the sampled population. Interestingly, multimodal SFS only occurs for very specific conditions, that *is* large immigration rates and intermediate duration of gene flow ($T_{GF} = T_{DIV}/100$, Figure 4d), and in those cases, the internal mode is only seen at high derived frequencies.

3.3 | Application to human data

All ten 1000G populations show clear uSFS at neutral sites (Figure 5). Among the four demographic scenarios tested on these human data (Supporting Information 5), only the scenario of genetic *isolation* fails to produce uSFS (Supporting Information 10), shows less good fit especially when looking at normalized SFS (Lapierre, Lambert, & Achaz, 2017), and a significantly lower likelihood than that of the three other scenarios (Supporting Information 11). Expected SFS is found very similar to the observed ones, and the estimated maximum likelihood values are found close to the maximum possible value (computed by assuming that the expected SFS entries would

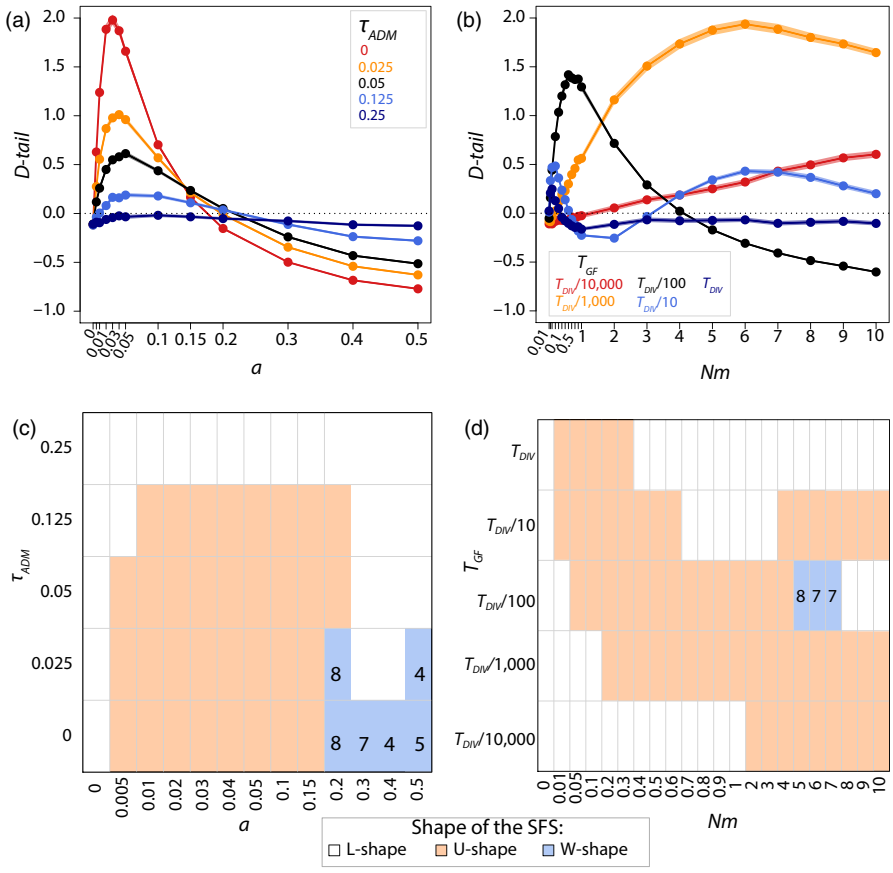

FIGURE 4 Effect of rate and age of gene flow on SFS properties, for $\tau_{DIV} = 2.5$ and n = 10 under an *IA* model with different admixture times (τ_{ADM}) and rates (a) (left panes) or under an *II* model with different gene flow onset (T_{GF}) and number of migrants per generation (Nm) (right panes). *D-tail* statistic (panes a and b) with dots and solid lines obtained from simulated data sets, and semi-transparent colours defining the 95% confidence intervals calculated from the block-bootstrap data sets (note that these confidence intervals are so small that they are barely visible on these figures); SFS shapes (panes c and d) with black numbers indicating the derived frequency i of the internal mode of W-shaped SFS

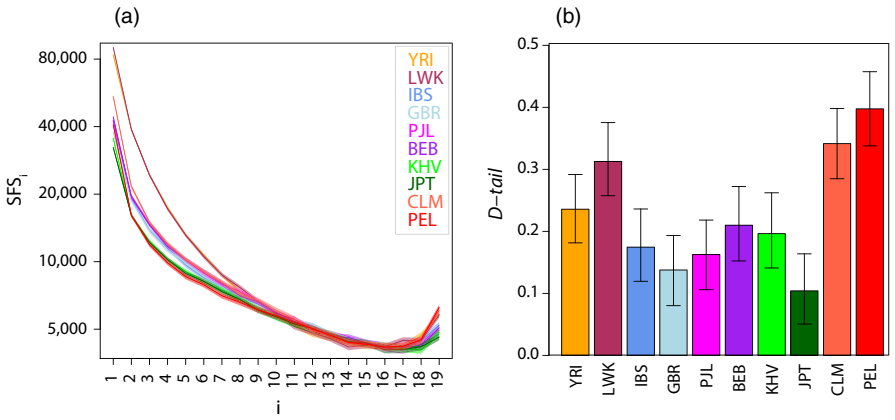

FIGURE 5 Neutral SFS (a) and associated *D-tail* statistics (b) observed in ten 1000G human samples. In pane a, SFS_i is the number of sites with a derived frequency *i*. In pane b, the whiskers indicate limits of 95% block-bootstrap confidence intervals

be equal to the observed SFS entries) for the three other scenarios: *isolation with ASM*, *admixture* and *immigration*. Therefore, we cannot distinguish which of these three scenarios is best on the sole basis of their likelihoods. However, we find that an average of 4.38% (2.75% – 7.59%) of ancestral state mis-assignment is necessary for the *isolation with ASM* model to fit the data. This value is one order of magnitude higher than that previously estimated in Yoruba (0.1%–0.3% in Lapierre, 2017) by using sites for which the nucleotide of an out-group species is different from the two nucleotides defining a SNP in a focal population (Baudry & Depaulis, 2003). It suggests that ASM in a context of genetic isolation is not the cause of the uSFS observed from the human neutral data, and that one of the two models involving gene flow is a more plausible explanation. Overall, the best parameters inferred from gene flow scenarios generally point to mild and recent gene flow (mean admixture rate $a = 0.06$ and time $T_{ADM} = 171$ generations ago for *admixture* scenario; on average, 72 migrants per generation for 540 generations for the *immigration* scenario, i.e. postlast glacial maximum for non-African populations).

4 | DISCUSSION

Gene flow is often overlooked as an explanation for the observation of an excess of high-frequency-derived alleles. Natural populations showing uSFS are usually considered as isolated but under selection (Li et al., 2012; Liu et al., 2017; Qanbari & Simianer, 2014; Sabeti et al., 2006). However, this strong assumption of genetic isolation is far from being warranted, as gene flow between populations seems to be the standard in non-human species (Sexton, Hangartner, & Hoffmann, 2014), sometimes even extending over species boundaries (Shurtliff, 2013; Wang et al., 2019) and persisting despite habitat fragmentation due to human activity (Corlatti, Hacklaänder, Frey-Roos, HacklÄnder, & Frey-Roos, 2009). For humans, numerous occurrences of gene flow between populations have been documented at every epochs and on every continent (Hellenthal et al., 2014). Human isolates rather seem to be an exception (Heutink, 2002) and seem to have emerged recently due to geographical and/or cultural barriers, for example populations living on islands or remote places (Roberts, 1976; Serre, Jakobi, & Babron, 1985), or being cultural minorities (Bideau, Brunet, Heyer, & Plauchu, 1994; Capocasa et al., 2013; Mourali-Chebil & Heyer, 2006), and usually present health and fitness issues (Charlesworth & Willis, 2009; Keller & Waller, 2002; Spielman, Brook, Briscoe, & Frankham, 2004). In this paper, the ten 1000G populations we study all show uSFS in their neutral fraction of genomes, where selection is supposed to have almost no effect (Pouyet et al., 2018), suggesting that they are not genetically isolated populations. More generally, we show with simulations that gene flow alone (i.e. in the absence of any selection, for two very contrasting models of gene flow) can actually easily produce an excess of high-frequency-derived alleles and uSFS. Interestingly, we find that uSFS can emerge from gene flow both by (a) the introduction of a few ancestral alleles (*II* and other *IA* models) and (b) by a massive input of derived alleles (during a partial genetic replacement, i.e. *IA* model with admixture rates larger than 0.5). This latter result extends previous ones (Wakeley & Aliacar, 2001), as not only mild gene flow can lead to an excess of high-frequency-derived alleles after a single admixture event. Furthermore, for higher rates of gene flow between deeply divergent populations, we manage to simulate W-shaped SFS, a signal that can also be produced by balancing selection (Bitarello et al., 2018; Croze, Živković, Stephan, & Hutter, 2016), associative overdominance (Gilbert, Pouyet, Excoffier, & Peischl, 2020) or in a heterogeneous structure resulting from divergent sources sampled as a single population (González-Martínez, Ridout, & Pannell, 2017).

Our results are in line with the fact that human populations are not genetically isolated, even though our study did not formally identify the source of recently incoming lineages. In our models, we used an unsampled or "ghost" population as the source of gene flow (Excoffier et al., 2013), which simply models a reservoir for some divergent lineages now found in the sampled population (Beerli, 2004; Slatkin, 2005). It can represent a population that separated a long time ago from the sampled population, as in the case of a secondary contact after a period of isolation, like in hybrid zones at the population or species level (Alcala, Jensen, Telenti, & Vuilleumier, 2016; Alcala & Vuilleumier, 2014; Hvala & Wood, 2012; Tine et al., 2014). Such hybridization events have occurred repeatedly in human evolution (e.g. between anatomically modern and archaic humans (Dannemann & Racimo, 2018)

or after long-distance dispersals between population of distinct ancestries (Fortes-Lima et al., 2018; Sedghifar, Brandvain, Ralph, & Coop, 2015; Verdu et al., 2014). Interestingly, if the source population is actually sampled, the joint SFS for the source and the target populations will reveal in the target population an excess of rare or event quite frequent derived alleles, for small and large immigration rates, respectively (Supporting Information 12), as previously reported in population or species having recently reconnected (Alcala et al., 2016; Alcala & Vuilleumier, 2014; Fraïsse et al., 2018; Tellier et al., 2011; Tine et al., 2014). Alternatively, as already mentioned above, the "ghost" population does not need to correspond to a real or an existing population, but can rather simply represent a set of populations surrounding the sampled population, as in large spatially structured populations, which can be described as a continent-island model (Excoffier, 2004; Hahn, 2018), for example like after a spatial expansion.

This last type of ghost (continent) population can be particularly relevant to model the history of human populations, as we were not able to identify the source of gene flow within the available 1,000 Genomes populations (Supporting Information 13). A consensus scenario for the worldwide expansion of humans is a serial founder effect out of Africa with limited archaic hybridization (Ramachandran et al., 2005; Stringer, 2014). As uSFS has actually been observed in simulations of range expansions, one could think that gene surfing having occurred during past human range expansions could explain the observed uSFS (Sousa et al., 2014). However, during human expansions, both recurrent founder effects at the front and migration between neighbouring demes in the wake of the front certainly occurred, such that gene surfing at the front could have promoted the fixation of different alleles in different sectors and a mixing of these sectors in the wake of the expansion could have led to uSFS (Peischl, Dupanloup, Bosshard, & Excoffier, 2016). We have run additional simulations to investigate the impact of gene flow during range expansions on the SFS (Supporting Information 14). We find that uSFS is only observed when gene flow between adjacent populations on the front is associated with the expansion, showing that gene surfing alone cannot lead to an excess of high-frequency-derived alleles. In addition, we find that uSFS can also stem from a Wahlund effect, that *is* when the SFS is computed from a population with hidden subdivisions (Supporting Information 15). Therefore, uSFS can emerge from naturally occurring gene flow or from artefactual structure resulting from the sampling of divergent lineages, as both will result in the potential mixing of differentially fixed alleles.

Whereas uSFS is never observed in completely isolated populations under a classical Kingman coalescent model, they can certainly exist under multiple-merger coalescent (MMC) models (Eldon et al., 2015; Pitman, 1999; Sagitov, 1999; Sargsyan & Wakeley, 2008; Schweinsberg, 2000; Tellier & Lemaire, 2014), which occur under recurrent episodes of selective sweeps or for extremely skewed distributions of offspring numbers (e.g. oyster, cod, bacteria and viruses (Árnason & Halldoŕsdóttir, 2015; Sargsyan & Wakeley, 2008; Tellier & Lemaire, 2014)). Since different parts of the genome can be differentially affected by selection, a mixture of classical and multiple-merger coalescent models could be used to model whole genomes (Rice et al., 2018). Contrastingly, gene flow into a population should affect the whole genome, even though effective migration rates may be affected by intragenomic selective processes as well (Petry, 1983; Sousa & Hey, 2013). It would therefore be interesting to include the effect of gene flow in the context of multiple-merger models as well. Along the same lines, procedure contrasting the SFS at different positions of the genome to evidence selection (Fay & Wu, 2000; Kim & Stephan, 2002; Nielsen et al., 2005; Pavlidis et al., 2010; Zeng, Fu, Shi, & Wu, 2006) or methods using the SFS to infer the distribution of fitness effects (Eyre-Walker & Keightley, 2007; Kim, Huber, & Lohmueller, 2017; Tataru, Mollion, Glémin, & Bataillon, 2017) do not take gene flow into account and could thus lead to biased inferences. We therefore hope that our study would promote the inclusion of gene flow when studying the effect of selection on genomic diversity.

ACKNOWLEDGEMENTS

The authors would like to thank Fanny Pouyet who provided us the 1000G neutral SFS; Stephan Peichel, Guillaume Achaz and Daniel Wegmann for helpful discussions. NM was partially supported by a Swiss National Science Foundation No 310030B_166605 to LE and by a Seal of Excellence Fund grant from the University of Berne (SELF2018-04).

CONFLICT OF INTEREST

None declared.

AUTHOR CONTRIBUTIONS

L.E. and N.M. designed the study and wrote the manuscript, L.E. performed *fastsimcoal2* parameter estimations from 1000Genomes data, N.M. run the simulations and analysed the results.

DATA AVAILABILITY STATEMENT

SFS from the simulations and for the ten studied 1,000 Genomes populations is available upon request to the authors. Furthermore, the input files necessary to repeat the simulations are provided in the Supplementary Information.

ORCID

Nina Marchi https://orcid.org/0000-0001-6624-5922

Laurent Excoffier https://orcid.org/0000-0002-7507-6494

REFERENCES

Akashi, H., & Schaeffer, S. W. (1997). Natural selection and the frequency distributions of "silent" DNA polymorphism in Drosophila. *Genetics*, *146*(1), 295–307.

Alcala, N., Jensen, J. D., Telenti, A., & Vuilleumier, S. (2016). The genomic signature of population reconnection following isolation: From theory to HIV. *G3: Genes, Genomes Genetics*, *6*(1), 107–120. https://doi.org/10.1534/g3.115.024208

AlcalaNicolas, Vuilleumier Séverine (2014). Turnover and accumulation of genetic diversity across large time-scale cycles of isolation and connection of populations. *Proceedings of the Royal Society B: Biological Sciences*, *281*, (1794), 20141369http://dx.doi.org/10.1098/rspb.2014.1369.

Andolfatto, P., & Przeworski, M. (2001). Regions of lower crossing over harbor more rare variants in African populations of Drosophila melanogaster. *Genetics*, *158*(2), 657–665.

ÁrnasonE., Halldórsdóttir K. (2015). Nucleotide variation and balancing selection at theCkmagene in Atlantic cod: analysis with multiple merger coalescent models. *PeerJ*, *3*, e786. http://dx.doi.org/10.7717/peerj.786

Baudry, E., & Depaulis, F. (2003). Effect of misoriented sites on neutrality tests with outgroup. *Genetics*, *165*(3), 1619–1622.

Beerli, P. (2004). Effect of unsampled populations on the estimation of population sizes and migration rates between sampled populations. *Molecular Ecology*, *13*(4), 827–836. https://doi.org/10.1111/j.1365-294X.2004.02101.x

Bideau, A., Brunet, G., Heyer, E., & Plauchu, H. (1994). La consanguinité, révélateur de la structure de la population. L'exemple de la vallée de la Valserine du XVIIIe siècle à nos jours. *Population (French Edition)*, *49*(1), 145–160.

Bitarello, B. D., De Filippo, C., Teixeira, J. C., Schmidt, J. M., Kleinert, P., Meyer, D., & Andres, A. M. (2018). Signatures of long-term balancing selection in human genomes. *Genome Biology and Evolution*, *10*(3), 939–955. https://doi.org/10.1093/gbe/evy054

Bustamante, C. D., Wakeley, J., Sawyer, S., & Hartl, D. L. (2001). Directional selection and the site-frequency spectrum. *Genetics*, *159*(4), 1779–1788.

Capocasa, M., Battaggia, C., Anagnostou, P., Montinaro, F., Boschi, I., Ferri, G., … Bisol, G. D. (2013). Detecting genetic isolation in human populations: A study of European language minorities. *PLoS One*, *8*(2), https://doi.org/10.1371/journal.pone.0056371

Charlesworth, D., & Willis, J. H. (2009). The genetics of inbreeding depression. *Nature Reviews Genetics*, *10*(11), 783–796. https://doi.org/10.1038/nrg2664

Cooper, B. S., Burrus, C. R., Ji, C., Hahn, M. W., & Montooth, K. L. (2015). Similar efficacies of selection shape mitochondrial and nuclear genes in both Drosophila melanogaster and Homo sapiens. *G3: Genes, Genomes, Genetics*, *5*(10), 2165–2176. https://doi.org/10.1534/g3.114.016493

Corlatti, L., Hacklaänder, K., Frey-Roos, F., HacklÄnder, K., & Frey-Roos, F. (2009). Ability of Wildlife Overpasses to Provide Connectivity and Prevent Genetic Isolation. *Conservation Biology*, *23*(3), 548–556. https://doi.org/10.1111/j.1523-1739.2008.01162.x

Croze, M., Živković, D., Stephan, W., & Hutter, S. (2016). Balancing selection on immunity genes: Review of the current literature and new analysis in Drosophila melanogaster. *Zoology*, *119*(4), 322–329. https://doi.org/10.1016/j.zool.2016.03.004

Cutter, A. D. (2019). *A Primer of Molecular Population Genetics*. Oxford: OUP Oxford. https://books.google.ch/books?id=_c6aDwAAQBAJ

Dannemann, M., & Racimo, F. (2018). Something old, something borrowed: Admixture and adaptation in human evolution. *Current Opinion in Genetics and Development*, *53*, 1–8. https://doi.org/10.1016/j.gde.2018.05.009

de Manuel, M., Kuhlwilm, M., Frandsen, P., Sousa, V. C., Desai, T., Prado-Martinez, J., … Marques-Bonet, T. (2016). Chimpanzee genomic diversity reveals ancient admixture with bonobos. *Science*, *354*(6311), 477–481. https://doi.org/10.1126/science.aag2602

Eldon, B., Birkner, M., Blath, J., & Freund, F. (2015). Can the site-frequency spectrum distinguish exponential population growth from multiple-merger Coalescents? *Genetics*, *199*(3), 841–856. https://doi.org/10.1534/genetics.114.173807

Ewens, W. J. (1972). The sampling theory of selectively neutral alleles. *Theoretical Population Biology*, *3*(1), 87–112. https://doi.org/10.1016/0040-5809(72)90035-4

Excoffier, L. (2004). Patterns of DNA sequence diversity and genetic structure after a range expansion: Lessons from the infinite-island model. *Molecular Ecology*, *13*(4), 853–864. https://doi.org/10.1046/j.1365-294X.2003.02004.x

Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V. C., & Foll, M. (2013). Robust demographic inference from genomic and SNP data. *PLoS Genetics*, *9*(10), e1003905. https://doi.org/10.1371/journal.pgen.1003905

Eyre-Walker, A., & Keightley, P. D. (2007). The distribution of fitness effects of new mutations. *Nature Reviews Genetics*, *8*(8), 610–618. https://doi.org/10.1038/nrg2146

Fay, J. C., & Wu, C. I. (2000). Hitchhiking under positive Darwinian selection. *Genetics*, *155*(3), 1405–1413. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1461156&tool=pmcentrez&rendertype=abstract

Fortes-Lima, C., Bybjerg-Grauholm, J., Marin-Padrón, L. C., Gomez-Cabezas, E. J., Bækvad-Hansen, M., Hansen, C. S., … Marcheco-Teruel, B. (2018). Exploring Cuba's population structure and demographic history using genome-wide data. *Scientific Reports*, *8*(1), 11422. https://doi.org/10.1038/s41598-018-29851-3

Fraïsse, C., Roux, C., Gagnaire, P. A., Romiguier, J., Faivre, N., Welch, J. J., & Bierne, N. (2018). The divergence history of European blue mussel species reconstructed from Approximate Bayesian Computation: the effects of sequencing techniques and sampling strategies. *PeerJ*, *6*, e5198. https://doi.org/10.7717/peerj.5198

Fu, Y. X. (1995). Statistical properties of segregating sites. *Theoretical Population Biology*, *48*(2), 172–197. https://doi.org/10.1006/tpbi.1995.1025

Garrigan, D., & Hammer, M. F. (2006). Reconstructing human origins in the genomic era. *Nature Reviews Genetics*, *7*(9), 669–680. https://doi.org/10.1038/nrg1941

Geneva, A., & Garrigan, D. (2010). Population genomics of secondary contact. *Genes*, *1*(1), 124–142. https://doi.org/10.3390/genes1010124

Gilbert, K. J., Pouyet, F., Excoffier, L., & Peischl, S. (2020). Transition from background selection to associative overdominance promotes diversity in regions of low recombination. *Current Biology*, *30*(1), 101–107.e3. https://doi.org/10.1016/j.cub.2019.11.063

González-Martínez, S. C., Ridout, K., & Pannell, J. R. (2017). Range expansion compromises adaptive evolution in an outcrossing plant. *Current Biology*, *27*(16), 2544–2551.e4. https://doi.org/10.1016/j.cub.2017.07.007

Hahn, M. W. (2018). *Molecular population genetics*. Sunderland, MA: Sinauer Associates. ISBN 978-0878939657.

Hellenthal, G., Busby, G. B. J., Band, G., Wilson, J. F., Capelli, C., Falush, D., & Myers, S. (2014). Europe PMC Funders Group Europe PMC Funders Author Manuscripts A genetic atlas of human admixture history. *Science*, *343*(6172), 747–751. https://doi.org/10.1126/science.1243518.A

Henn, B. M., Botigué, L. R., Peischl, S., Dupanloup, I., Lipatov, M., Maples, B. K., … Bustamante, C. D. (2015). Distance from sub-Saharan Africa predicts mutational load in diverse human genomes. *Proceedings of the National Academy of Sciences*, *113*(4), E440–E449. https://doi.org/10.1073/pnas.1510805112

Hernandez, R. D., Williamson, S. H., & Bustamante, C. D. (2007). Context dependence, ancestral misidentification, and spurious signatures of natural selection. *Molecular Biology and Evolution*, *24*(8), 1792–1800. https://doi.org/10.1093/molbev/msm108

Heutink, P. (2002). Gene finding in genetically isolated populations. *Human Molecular Genetics*, *11*(20), 2507–2515. https://doi.org/10.1093/hmg/11.20.2507

Hudson, R. R., & Coyne, J. A. (2002). Mathematical Consequences of the Genealogical Species Concept. *Evolution*, *56*(8), 1557. https://doi.org/10.1554/0014-3820(2002)056[1557:mcotgs]2.0.co;2

Huerta-Sanchez, E., Durrett, R., & Bustamante, C. D. (2008). Population genetics of polymorphism and divergence under fluctuating selection. *Genetics*, *178*(1), 325–337. https://doi.org/10.1534/genetics.107.073361

Hvala, J. A., & Wood, T. E. (2012). Speciation: Introduction. In: *eLS*, (Issue July). Chichester, UK:John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470015902.a0001709.pub3

Keller, L., & Waller, D. M. (2002). Inbreeding effects in wild populations. *Trends in Ecology & Evolution*, *17*(5), 230–241. https://doi.org/10.1016/S0169-5347(02)02489-8

Kim, B. Y., Huber, C. D., & Lohmueller, K. E. (2017). Inference of the distribution of selection coefficients for new nonsynonymous mutations using large samples. *Genetics*, *206*(1), 345–361. https://doi.org/10.1534/genetics.116.197145

Kim, Y., & Stephan, W. (2000). Joint effects of genetic hitchhiking and background selection on neural variation. *Genetics*, *155*(3), 1415–1427.

Kim, Y., & Stephan, W. (2002). Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics*, *160*(2), 765–777.

Lapierre, M. (2017). Extensions du modèle standard neutre pertinentes pour l'analyse de la diversité génétique. http://www.theses.fr/2017PA066395/document

Lapierre, M., Blin, C., Lambert, A., Achaz, G., & Rocha, E. P. C. (2016). The impact of selection, gene conversion, and biased sampling on the assessment of microbial demography. *Molecular Biology and Evolution*, *33*(7), 1711–1725. https://doi.org/10.1093/molbev/msw048

Lapierre, M., Lambert, A., & Achaz, G. (2017). Accuracy of demographic inferences from the site frequency spectrum: The case of the yoruba population. *Genetics*, *206*(1), 139–449. https://doi.org/10.1534/genetics.116.192708

Li, J., Li, H., Jakobsson, M., Li, S., SjÖdin, P., & Lascoux, M. (2012). Joint analysis of demography and selection in population genetics: Where do we stand and where could we go? *Molecular Ecology*, *21*(1), 28–44. https://doi.org/10.1111/j.1365-294X.2011.05308.x

Liu, Q., Zhou, Y., Morrell, P. L., Gaut, B. S., & Ge, S. (2017). Deleterious variants in Asian Rice and the potential cost of domestication. *Molecular Biology and Evolution*, *34*(4), 908–924. https://doi.org/10.1093/molbev/msw296

Malaspinas, A.-S., Westaway, M. C., Muller, C., Sousa, V. C., Lao, O., Alves, I., … Willerslev, E. (2016). A genomic history of Aboriginal Australia. *Nature*, *538*(7624), 207–214. https://doi.org/10.1038/nature18299

Morton, B. R., Dar, V. U. N., & Wright, S. I. (2009). Analysis of site frequency spectra from Arabidopsis with context-dependent corrections for ancestral misinference. *Plant Physiology*, *149*(2), 616–624. https://doi.org/10.1104/pp.108.127787

Mourali-Chebil, S., & Heyer, E. (2006). Evolution of inbreeding coefficients and effective size in the population of Saguenay Lac-St.-Jean (Quebec). *Human Biology*, *78*(4), 495–508. https://doi.org/10.1353/hub.2006.0056

Murray, C., Huerta-Sanchez, E., Casey, F., & Bradley, D. G. (2010). Cattle demographic history modelled from autosomal sequence variation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *365*(1552), 2531–2539. https://doi.org/10.1098/rstb.2010.0103

Nielsen, R., Williamson, S., Kim, Y., Hubisz, M. J., Clark, A. G., & Bustamante, C. (2005). Genomic scans for selective sweeps using SNP data. *Genome Research*, *15*(11), 1566–1575. https://doi.org/10.1101/gr.4252305

Pagani, L., Schiffels, S., Gurdasani, D., Danecek, P., Scally, A., Chen, Y., … Tyler-Smith, C. (2015). Tracing the Route of Modern Humans out of Africa by Using 225 Human Genome Sequences from Ethiopians and Egyptians. *The American Journal of Human Genetics*, *96*(6), 986–991. https://doi.org/10.1016/j.ajhg.2015.04.019

Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., … Reich, D. (2012). Ancient Admixture in Human History. *Genetics*, *192*(3), 1065–1093. https://doi.org/10.1534/genetics.112.145037

Pavlidis, P., Jensen, J. D., & Stephan, W. (2010). Searching for footprints of positive selection in whole-genome SNP data from nonequilibrium populations. *Genetics*, *185*(3), 907–922. https://doi.org/10.1534/genetics.110.116459

Peischl, S., Dupanloup, I., Bosshard, L., & Excoffier, L. (2016). Genetic surfing in human populations: From genes to genomes. *Current Opinion in Genetics and Development*, *41*, 53–61. https://doi.org/10.1016/j.gde.2016.08.003

Petry, D. (1983). The effect on neutral gene flow of selection at a linked locus. *Theoretical Population Biology*, *23*(3), 300–313. https://doi.org/10.1016/0040-5809(83)90020-5

Pitman, J. (1999). Coalescents with multiple collisions. *The Annals of Probability*, *27*(4), 1870–1902. https://doi.org/10.1214/aop/1022874819

Pouyet, F., Aeschbacher, S., Thiéry, A., & Excoffier, L. (2018). Background selection and biased gene conversion affect more than 95% of the human genome and bias demographic inferences. *Elife*, *7*, 1–25. https://doi.org/10.7554/eLife.36317

Price, N., Moyers, B. T., Lopez, L., Lasky, J. R., Grey Monroe, J., Mullen, J. L., … McKay, J. K. (2018). Combining population genomics and fitness QTLs to identify the genetics of local adaptation in Arabidopsis thaliana. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(19), 5028–5033. https://doi.org/10.1073/pnas.1719998115

Przeworski, M. (2002). The signature of positive selection at randomly chosen loci. *Genetics*, *160*(3), 1179–1189.

Qanbari, S., & Simianer, H. (2014). Mapping signatures of positive selection in the genome of livestock. *Livestock Science*, *166*(1), 133–143. https://doi.org/10.1016/j.livsci.2014.05.003

Raghavan, M., Steinrucken, M., Harris, K., Schiffels, S., Rasmussen, S., DeGiorgio, M., … Willerslev, E. (2015). Genomic evidence for the Pleistocene and recent population history of Native Americans. *Science*, *349*(6250), aab3884–aab3884. https://doi.org/10.1126/science.aab3884

Ramachandran, S., Deshpande, O., Roseman, C. C., Rosenberg, N. A., Feldman, M. W., & Cavalli-Sforza, L. L. (2005). Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(44), 15942–15947. https://doi.org/10.1073/pnas.0507611102

Rice, D. P., Novembre, J., & Desai, M. M. (2018). Distinguishing multiple-merger from Kingman coalescence using two-site frequency spectra. *BioRxiv Preprint*, 461517. https://doi.org/10.1101/461517

Roberts, D.-F. (1976). Les concepts d'isolats. *L'étude des isolats. Limites et espoirs* (pp. 75–92). Paris: INED.

Ronen, R., Udpa, N., Halperin, E., & Bafna, V. (2013). Learning natural selection from the site frequency spectrum. *Genetics*, *195*(1), 181–193. https://doi.org/10.1534/genetics.113.152587

Sabeti, P. C. (2006). Positive natural selection in the human lineage. *Science*, *312*(5780), 1614–1620. https://doi.org/10.1126/science.1124309

Sagitov, S. (1999). The general coalescent with asynchronous mergers of ancestral lines. *Journal of Applied Probability*, *36*(4), 1116–1125. https://doi.org/10.1239/jap/1032374759

Sargsyan, O., & Wakeley, J. (2008). A coalescent process with simultaneous multiple mergers for approximating the gene genealogies of many marine organisms. *Theoretical Population Biology*, *74*(1), 104–114. https://doi.org/10.1016/j.tpb.2008.04.009

Scally, A., & Durbin, R. (2012). Revising the human mutation rate: Implications for understanding human evolution. *Nature Reviews Genetics*, *13*(11), 745–753. https://doi.org/10.1038/nrg3295

Schiffels, S., & Durbin, R. (2014). Inferring human population size and separation history from multiple genome sequences. *Nature Genetics*, *46*(8), 919–925. https://doi.org/10.1038/ng.3015

SchweinsbergJason (2000). Coalescents with Simultaneous Multiple Collisions. *Electronic Journal of Probability*, *5*, 1–50. http://dx.doi.org/10.1214/ejp.v5-68

Sedghifar, A., Brandvain, Y., Ralph, P., & Coop, G. (2015). The spatial mixing of genomes in secondary contact zones. *Genetics*, *201*(1), 243–261. https://doi.org/10.1534/genetics.115.179838

Serre, J. L., Jakobi, L., & Babron, M. (1985). A genetic isolate in the French Pyrenees: Probabilities of origin of genes and inbreeding. *Journal of Biosocial Science*, *17*, 405–414. https://doi.org/10.1017/S0021932000015923

Sexton, J. P., Hangartner, S. B., & Hoffmann, A. A. (2014). Genetic isolation by environment or distance: Which pattern of gene flow is most common? *Evolution*, *68*(1), 1–15. https://doi.org/10.1111/evo.12258

Shurtliff, Q. R. (2013). Mammalian hybrid zones: A review. *Mammal Review*, *43*(1), 1–21. https://doi.org/10.1111/j.1365-2907.2011.00205.x

Sikora, M., Pitulko, V. V., Sousa, V. C., Allentoft, M. E., Vinner, L., Rasmussen, S., … Willerslev, E. (2019). The population history of northeastern Siberia since the Pleistocene. *Nature*, *570*(7760), 182–188. https://doi.org/10.1038/s41586-019-1279-z

Sikora, M., Seguin-Orlando, A., Sousa, V. C., Albrechtsen, A., Korneliussen, T., Ko, A., … Willerslev, E. (2017). Ancient genomes show social and reproductive behavior of early Upper Paleolithic foragers. *Science*, *358*(6363), 659–662. https://doi.org/10.1126/science.aao1807

Slatkin, M. (2005). Seeing ghosts: The effect of unsampled populations on migration rates estimated for sampled populations. *Molecular Ecology*, *14*(1), 67–73. https://doi.org/10.1111/j.1365-294X.2004.02393.x

Sousa, V., & Hey, J. (2013). Understanding the origin of species with genome-scale data: Modelling gene flow. *Nature Reviews Genetics*, *14*(6), 404–414. https://doi.org/10.1038/nrg3446

Sousa, V., Peischl, S., & Excoffier, L. (2014). Impact of range expansions on current human genomic diversity. *Current Opinion in Genetics and Development*, *29*, 22–30. https://doi.org/10.1016/j.gde.2014.07.007

Spence, J. P., & Song, Y. S. (2019). Inference and analysis of population-specific fine-scale recombination maps across 26 diverse human populations. *Science*. *Advances*, *5*(10), eaaw9206. https://doi.org/10.1126/sciadv.aaw9206

Spielman, D., Brook, B. W., Briscoe, D., & a, & Frankham, R., (2004). Does inbreeding and loss of genetic diversity reduce disease resistance? *Conservation Genetics*, *5*, 439–448. https://doi.org/10.1023/B:COGE.0000041030.76598.cd

Steinrücken, M., Kamm, J., Spence, J. P., & Song, Y. S. (2019). Inference of complex population histories using whole-genome sequences from multiple populations. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(34), 17115–17120. https://doi.org/10.1073/pnas.1905060116

Stephan, W. (2016). Signatures of positive selection: From selective sweeps at individual loci to subtle allele frequency changes in polygenic adaptation. *Molecular Ecology*, *25*(1), 79–88. https://doi.org/10.1111/mec.13288

Stringer, C. (2014). Why we are not all multiregionalists now. *Trends in Ecology and Evolution*, *29*(5), 248–251. https://doi.org/10.1016/j.tree.2014.03.001

Tataru, P., Mollion, M., Glémin, S., & Bataillon, T. (2017). Inference of distribution of fitness effects and proportion of adaptive substitutions from polymorphism data. *Genetics*, *207*(November), 1103–1119. https://doi.org/10.1534/genetics.117.300323/-/DC1.1

Tellier, A., & Lemaire, C. (2014). Coalescence 2.0: A multiple branching of recent theoretical developments and their applications. *Molecular Ecology*, *23*(11), 2637–2652. https://doi.org/10.1111/mec.12755

Tellier, A., Pfaffelhuber, P., Haubold, B., Naduvilezhath, L., Rose, L. E., Städler, T., … Metzler, D. (2011). Estimating parameters of speciation models based on refined summaries of the joint site-frequency spectrum. *PLoS One*, *6*(5), e18155. https://doi.org/10.1371/journal.pone.0018155

Tine, M., Kuhl, H., Gagnaire, P.-A., Louro, B., Desmarais, E., Martins, R. S. T., … Reinhardt, R. (2014). European sea bass genome and its variation provide insights into adaptation to euryhalinity and speciation. *Nature Communications*, *5*(May), 5770. https://doi.org/10.1038/ncomms6770

Venn, O., Turner, I., Mathieson, I., De Groot, N., Bontrop, R., & McVean, G. (2014). Strong male bias drives germline mutation in chimpanzees. *Science*, *344*(6189), 1272–1275. https://doi.org/10.1126/science.344.6189.1272

Verdu, P., Pemberton, T. J., Laurent, R., Kemp, B. M., Gonzalez-Oliver, A., Gorodezky, C., … a, & Malhi, R. S., (2014). Patterns of admixture and population structure in native populations of Northwest North America. *PLoS Genetics*, *10*(8), 1–17. https://doi.org/10.1371/journal.pgen.1004530

Wakeley, J. (1999). Nonequilibrium migration in human history. *Genetics*, *153*(4), 1863–1871. https://doi.org/10.1021/bk-1997-0664.ch001

Wakeley, J. (2000). The effects of subdivision on the genetic divergence of populations and species. *Evolution*, *54*(4), 1092–1101. https://doi.org/10.1111/j.0014-3820.2000.tb00545.x

Wakeley, J., & Aliacar, N. (2001). Gene genealogies in a metapopulation. *Genetics*, *159*(1997), 893–905.

Wang, X., Que, P., Heckel, G., Hu, J., Zhang, X., Chiang, C. Y., … Liu, Y. (2019). Genetic, phenotypic and ecological differentiation suggests incipient speciation in two Charadrius plovers along the Chinese coast. *BMC Evolutionary Biology*, *19*(1), 1–18. https://doi.org/10.1186/s12862-019-1449-5

Wright, S. (1931). Evolution in Mendelian populations. *Genetics*, *16*(2), 97–159. https://doi.org/10.1007/BF02459575

Zeng, K., Fu, Y. X., Shi, S., & Wu, C. I. (2006). Statistical tests for detecting positive selection by utilizing high-frequency variants. *Genetics*, *174*(3), 1431–1439. https://doi.org/10.1534/genetics.106.061432

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: MarchiN, Excoffier L. Gene flow as a simple cause for an excess of high-frequency-derived alleles. *Evol Appl*. 2020;00:1–10. https://doi.org/10.1111/eva.12998